DOCUMENT RESUME

ED 088 952 TM 003 519

AUTHOR Kohr, Richard L.; Games, Paul A.
TITLE [Procedures for Testing Differences Between Means in the Presence of Unequal N's and Variances].
PUB DATE Apr 74
NOTE 19p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, Illinois, April, 1974)

EDRS PRICE MF-$0.75 HC-$1.50
DESCRIPTORS *Hypothesis Testing; Probability Theory; Sampling; *Statistical Analysis; *Tests of Significance
IDENTIFIERS *Variance (Statistical)

ABSTRACT

An empirical sampling study investigated six procedures for testing differences between means in the presence of unequal n's and variances. Support was obtained for previcus research which found t robust to heterogeneous variances only when n's are equal and of moderate size. The procedure which emerged as providing the best control over Type I errors while maintaining satisfactory power in all test conditions was the Behrens-Fisher v statistic with Welch's solution for degrees of freedom (df). The general recommendations when the population variances are unknown are: (1) when n's are the same and equal to or greater than 20 it is permissible to use the t statistic with df = 2n - 2, but when n is less than 20 use v with Welch's solution for df; (2) when n's are unequal use the v statistic with the Welch adjustment for df. (Author)

PROCEDURES FOR TESTING $\mu_1 = \mu_2$ WITH UNEQUAL N's AND VARIANCES

Richard L. Kohr

Pennsylvania Department of Education

and

Paul A. Games

The Pennsylvania State University

Presented at the Annual Meeting of the

American Educational Research Association

Chicago, Illinois

April, 1974

Procedures for Testing $\mu_1 = \mu_2$ with Unequal n's and Variances

Richard L. Kohr
Pennsylvania Department of Education

and

Paul A. Games
The Pennsylvania State University

An empirical sampling study investigated six procedures for testing $\mu_1 = \mu_2$ in the presence of unequal n's and variances. Support was obtained for previous research which found t robust to heterogeneous variances only when n's are equal and of moderate size. The procedure which emerged as providing the best control over Type I errors while maintaining satisfactory power in all test conditions was the Behrens-Fisher v statistic with Welch's solution for df. The general recommendations when the population variances are unknown are: (1) when n's are equal and $\geq$ 20 it is permissible to use the t statistic with df = 2n - 2, but when $n < 20$ use v with Welch's solution for df; (2) when n's are unequal use the v statistic with the Welch adjustment for df.

Procedures for Testing $\mu_1 = \mu_2$ with Unequal n's and Variances

Richard L. Kohr
Pennsylvania Department of Education

and

Paul A. Games
The Pennsylvania State University

Scheffé (1970, p. 1501) begins a recent article by stating "The most frequently occurring problem in applied statistics is, in my opinion, the comparison of the means of two populations,... .Let σ_1^2 and σ_2^2 denote the population variances. It is called the Behrens-Fisher problem if the ratio $\theta = \sigma_1^2/\sigma_2^2$ is unknown and the assumption is added that the populations are normal. The normality assumption is of no practical importance for any of the solutions... based mainly on the difference of the sample means are robust against its violation." Despite the fact that θ is unknown in most empirical studies, the Behrens-Fisher problem is ignored in many behavioral statistics books (Ferguson, 1966; Guilford, 1965; Glass & Stanley, 1970; Dayton, 1970; Myers, 1966; Watt & Bridges, 1967). Instead the null hypothesis $\mu_1 = \mu_2$ is usually tested by the conventional $\underline{t}$ test with $df = n_1 + n_2 - 2$ as if the assumption that $\theta = 1.0$ were based on something other than the experimenter's hope. This article will attempt to point out that there are other practical solutions to this problem which do not need this dubious assumption.

Behrens (1929) provided the first "exact" solution for this problem, which coincides with the Bayesian solution of Jeffreys (1940) and Savage (1961). Fisher (1935) extended Behrens' work and the statistic used became known as the Behrens-Fisher statistic, $\underline{v}$ (Winer, 1962, p. 37). Tabled probability values were prepared by Sukhatme (1938) and are reproduced in Fisher and Yates (1963).

$$\underline{v} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

Welch (1947) proposed an exact test for $\underline{v}$, which was subsequently tabled by Aspin (1949). Welch (1949) reports little difference between these critical values and an approximate solution obtained by adjusting the df of $\underline{t}$ (Winer, 1962, p. 37; Kirk, 1968, p. 98) so that,

$$df = \frac{(df_1)(df_2)}{df_2 C + df_1(1 - C)^2}, \text{ where } C = \frac{s_1^2/n_1}{s_1^2/n_1 + s_2^2/n_2}.$$

The critical value of $\underline{t}$ (α, df) is contrasted with $\underline{v}$ to complete the test. We shall label this the $\underline{v}$-$\underline{W}$ solution.

Several other methods have also been suggested which seek to approximate the sampling distribution of $\underline{v}$. Satterthwaite, (1946, p. 114) offered a df adjustment where df can range from $n_s - 1$ (where n_s is the smaller n) to $n_1 + n_2 - 2$. The df solutions of Welch and Satterthwaite are algebraically equivalent. Dixon and Massey (1957, p. 124) and Hays (1963, p. 322) present a formula which is a slight variation of the one given by Satterthwaite. A critical $\underline{t}$ procedure on the conservative side was devised by Cochran and Cox (1950, p. 92).

Scheffé (1943) described the following solution to the Behrens-Fisher problem. Let group 1 and 2 represent the smaller and larger groups respectively. The observations in each group are randomly ordered, and n_1 values of a new variate, U_i, are computed by:

$$U_i = X_{1i} - (n_1/n_2)X_{2i}, \text{ where } i = 1, \ldots n_1$$

A t-test that the mean of the U variate is 0.0 (or a corresponding confidence interval) is then conducted using the U's as the input data. Scheffé demonstrated that the expected value of the width of 95% confidence intervals produced by the above method was no greater than 11% longer than the width when the population variances were known and $n \geq 10$. Some statisticians are unhappy with randomization procedures such as the above, since E's may obtain different results with the same set of initial data.

Scheffé himself suggested discarding this method because he apparently found that some experimenters who did not like the width of the confidence interval obtained merely rerandomized and computed a second interval (Scheffé, 1970, footnote 3). However, if the data is entered into a computer, with automation doing the randomization, the experimenter usually will be oblivious of this temptation.

Another possible method is a chain logic approach. Games and Klare (1967, p. 494) suggested using t if equal n's greater than 10 are present. Otherwise $F = s_1^2/s_2^2$ is computed and tested at $\alpha = .10$. If this test is significant, v with Welch's df is used. If not, the conventional t is used. The probability of Type I errors, P(EI), resulting from this method may be expected to fall between that of t and of v with Welch's adjustment to df.

There are many other approximations to the sampling distribution of v, and many other procedures for handling the Behrens-Fisher problem. A search of the mathematical statistics literature obtained over 50 references of which 27 were published in 1960 or later.

Scheffé (1970) compared six solutions: the Behrens-Fisher solution; his 1943 solution, S; the conventional t test; the use of v with $t(\alpha, n_s - 1)$; the Welch-Aspin solution; and the v-W. His conclusions on the conventional t test may shock naive users. "....this solution of the Behrens-Fisher problem

is asymptotically incorrect for large unequal sample sizes, elementary calculations showing that $\beta(\theta)$ may take on any value between 0 and 1 for any $\alpha (0 < \alpha < 1)$ and for suitable θ and suitably large n_1, n_2. The practical conclusion is that this solution should never be used unless n_1 and n_2 are equal or nearly so (1970, p. 1506)." His $\beta(\theta)$ = P(EI) in the present paper. When $n_1 = n_2 = n$ then the limit of P(EI) is $P\left[|\underline{t}(n - 1)|\right] > \underline{t}(\alpha, 2n - 2)$. Thus when $\alpha = .05$ and n = 10, the limit of P(EI) is .065. For larger equal n's, P(EI) will deviate less from α.

Scheffé (1970) concludes that only the use of $\underline{v}$ with $\underline{t}(\alpha, n_s - 1)$ as the critical value, the Welch-Aspin solution, and the $\underline{v}$-$\underline{W}$ solution are practical. He finds little difference between the last two. The use of $\underline{v}$ with $\underline{t}$ $(\alpha, n_s - 1)$ is conservative and has a lower power than the $\underline{v}$-$\underline{W}$. Thus he concludes "....Welch's approximate $\underline{t}$-solution, which requires only the ubiquitous $\underline{t}$-tables, is a satisfactory practical solution of the Behrens-Fisher problem" (1970, p. 1505).

Wang (1971) presents results that confirm the similarity of the Welch-Aspin and $\underline{v}$-$\underline{W}$ solutions, and the excellence of control of P(EI) by the $\underline{v}$-$\underline{W}$ test. With n_s as small as 5 and θ varying from .0078 to 128, he finds that P(EI) does not deviate from α by more than .0035. When n_s is increased to 7, the maximum deviation drops to .0018; with n = 11, to .0005; and to .0002 with larger n_s values. Thus the $\underline{v}$-$\underline{W}$ test is supported as an excellent procedure whenever the population variances are unknown, and as an absolute necessity when the sample sizes are unequal.

Method

An orthogonal design was used to contrast the P(EI) and power of six methods across three dimensions. The condition of heterogeneity of variance

was represented by seven levels of variance ratios (VR): σ_1^2/σ_2^2 of .025, .25, .5, 1.0, 2.0, 4.0, and 40.0. The proportionality of n's was represented by four levels of sample size ratio's (NR): n_2/n_1 = 1.00, 1.25, 1.50, and 3.00. Sample sizes (SS) was represented by three levels: n_1 values of 4, 12, and 24. These three dimensions were completely crossed making 7 x 4 x 3 = 84 conditions that were investigated in the study. Six points on the power curve were established for each condition. This always included the condition where $\mu_1 - \mu_2 = 0.0$ (i.e., the null is true), plus five other evenly spaced points chosen to avoid a ceiling effect for the last condition. The six procedures used were: (1) t referred to the t distribution with df = $n_1 + n_2 - 2$; (2 and 3) v referred to the t distribution with Welch's (v-W) and Dixon and Massey's (v-DM) solutions for df; (4) v referred to Cochran and Cox's critical t (v-CC); (5) chain logic approach of Games and Klare (G & K); (6) Scheffé's procedure (S).

A FORTRAN IV computer program written for the IBM 360/67 generated a population of 9998 cases having the following parameters: $\mu = 0.0012$; $\sigma^2 = 1.0129$; skewness, $\gamma_1 = 0.002$; kurtosis, $\gamma_2 = -0.0322$. Each computer run involved the following steps:

(1) Generation of the above population.

(2) A sample of n_1 cases (representing group one) was drawn from the population. All sampling was with replacement. The normal deviate was multiplied by a constant representing the desired standard deviation to produce different variances when desired. A constant (0 to 11.0) was added to the randomly drawn value so that the population mean of group one might differ from that of group two when desired

(3) A sample of n_2 cases (representing group two) was drawn from the

population. The population mean of group two was fixed at zero throughout the study. Since the mean for group one was the only one to vary, the true difference, $\mu_1 - \mu_2$ was always positive in value. The variance of group two was manipulated in the same manner as in group one.

(4) The sample statistics required for conducting all significance tests for each of the six procedures were computed.

(5) Significance tests were conducted for all six statistical procedures at the .02 and .05 levels and the number of significant results were tabulated for each.

(6) Steps (2) through (5) were repeated until 250 simulated experiments had been obtained.

(7) The proportion of significant results at each significance level for each statistical procedure were punched.

(8) Steps (2) through (7) were repeated until four blocks of 250 samples each were drawn.

(9) Steps (2) through (8) were repeated for each value of $\mu_1 - \mu_2$.

(10) Steps (2) through (9) were repeated if $n \neq n$ and $\sigma_1^2 \neq \sigma_2^2$, switching the desired variance from group two to group one. This permitted the larger variance to be combined with both the larger and the smaller n.

All calculations were performed in double precision to insure the greatest possible accuracy. The pseudo-random number generator used in the study was prepared by Knoble (1969). Single precision, floating point values are returned which are approximately serially independent and uniformly distributed on the unit interval. The cycle length is $2^{31} - 2$. Different sequences of pseudo-random numbers are obtained by entering the sequence at a different point. Statistical properties of the generator may be found in Payne, Rabung, and Bogyo (1969) and Lewis, Goodman, and Miller (1969).

RESULTS

Case Where the Null Hypothesis is True

For brevity, tabular results are presented for only four of the techniques. The critical values of Cochran and Cox are greater than or equal to those of the Welch approximation. The Welch critical values are greater than or equal to those of Dixon and Massey. Since methods 2, 3, and 4 consisted in referring $\underline{v}$ to these critical values, it is clear that P(EI) and power will vary systematically between them. The study confirmed that the $\underline{v}$-$\underline{W}$ solution P(EI)'s are closest to alpha while the v-CC method is systematically conservative, and the $\underline{v}$-DM method is systematically permissive.

Table 1 summarizes P(EI) of the $\underline{v}$-$\underline{W}$ and the remaining competitive solutions for the conditions producing the greatest ($n_1 = 4$) and least ($n_2 = 24$) discrepancies. Since there are 1000 observations, any proportion less than .0365 and greater than .0635 is significantly different from alpha at the .05 level.

In general, the $\underline{t}$ statistic revealed the expected distortion to P(EI) when both variances and n's are unequal. Conservative results occur when the largest variance is combined with the larger sample, while an excessive number of rejections of H_o occur when the largest variance is paired with the smaller sample. As the n-ratio increases this effect also increases. It is clear that the $\underline{t}$ test is <u>not</u> robust to the homogeneity of variance assumption when n's are unequal. Note that increasing the sample size does <u>not</u> eliminate the distortion to P(EI) when $\underline{t}$ is used with unequal n's. The Welch, Games and Klare, and Scheffé techniques all reduced the $\underline{t}$'s fluctuations, the improvement in control over P(EI) generally increasing with increased sample sizes. Only slight deviations from the theoretical .05 value occurred for these three methods.

TABLE 1

PROPORTION OF SIGNIFICANT RESULTS
WHEN H_0 IS TRUE, $\alpha = .05$

		Procedures							
		Small Sample Size				Large Sample Size			
N-R	V-R	t	v-W	G&K	S	t	t-W	G&K	S
1.0	0.025	.092*	.065*	.068*	.057	.060	.058	.060	.056
1.0	0.25	.060	.046	.051	.049	.060	.058	.060	.056
1.0	0.5	.061	.049	.051	.049	.059	.056	.059	.057
1.0	1.0	.053	.047	.052	.051	.048	.048	.048	.046
1.0	2.0	.048	.040	.046	.042	.038	.038	.038	.035*
1.0	4.0	.056	.047	.053	.057	.049	.047	.049	.044
1.0	40.0	.088*	.060	.062	.054	.053	.052	.053	.052
1.25	0.025	.062	.062	.062	.061	.034*	.046	.046	.051
1.25	0.25	.042	.042	.042	.030*	.030*	.045	.045	.046
1.25	0.5	.036*	.037	.036*	.047	.057	.064*	.063	.060
1.25	1.0	.055	.050	.051	.048	.055	.056	.057	.058
1.25	2.0	.066*	.051	.056	.044	.062	.047	.054	.052
1.25	4.0	.081*	.060	.069*	.050	.059	.048	.048	.046
1.25	40.0	.104*	.057	.058	.053	.092*	.046	.046	.047
1.5	0.025	.031*	.048	.046	.061	.015*	.046	.046	.046
1.5	0.25	.033*	.039	.038	.045	.036*	.051	.050	.053
1.5	0.5	.034*	.040	.039	.050	.029*	.047	.043	.046
1.5	1.0	.043	.039	.043	.051	.047	.048	.048	.049
1.5	2.0	.062	.048	.053	.045	.073*	.054	.058	.052
1.5	4.0	.094*	.063	.078*	.048	.085*	.048	.048	.049
1.5	40.0	.156*	.060	.062	.052	.115*	.050	.050	.047
3.0	0.025	.005*	.053	.053	.049	.000*	.041	.041	.048
3.0	0.25	.005*	.043	.033	.056	.009*	.042	.041	.057
3.0	0.5	.028*	.060	.056	.056	.018*	.047	.043	.063
3.0	1.0	.051	.061	.067*	.065*	.055	.058	.053	.063
3.0	2.0	.104*	.053	.091*	.044	.096*	.051	.059	.057
3.0	4.0	.145*	.059	.102*	.045	.144*	.046	.046	.048
3.0	40.0	.301*	.057	.063	.058	.221*	.046	.046	.057

* Represents significant deviation from α

Special Case of Equal Sample Sizes

Equal sample sizes represent a special case in which prior research has revealed general robustness of t in the face of heterogeneous variances. Sample size, variance ratio, and the procedures of Table 1 were factors in a 3 factor ANOVA of P(EI).

The major significant source of variance was the procedures main effect which accounted for an estimated 18 per cent of the variance. The mean P(EI) for the t, Welch, Games and Klare, and Scheffé procedures were .0576, .0511, .0540, .0508 respectively. For the intermediate and large equal n cases all four procedures exercise adequate control over Type I errors. The Welch and Scheffé methods were the most stable across situations.

Case of Varying Degrees of Deviation from Ho (Power)

For a more complete comparison of the four procedures a series of power curves were plotted. The power curves that are graphed and discussed in the following sections are those of the intermediate n case. The essential difference between the curves of the intermediate n case and those of the small n case is that the curves for the Welch, Scheffé, and Games and Klare procedures are further apart when n's are small and become progressively closer as sample size increases.

Special Case of Equal Variances

When $\sigma_1^2 = \sigma_2^2$ and the populations are normally distributed, as in this study, then the t test is the most powerful possible test. When the variance ratio was 1.0 and the n-ratio was 1.25 ($n_1 = 12$, $n_2 = 15$) the power curve for t was indistinguishable from those for Welch's solution and the Games and Klare procedure. As the n-ratio increased to 3.0 ($n_1 = 12$, $n_2 = 36$) some separation of the power curves occurred. In this region the t demonstrated the highest power

with the Welch solution close behind followed by the Scheffé procedure. The power loss by using the Welch technique did not exceed .05.

Unequal Variances with Unequal n's

When the larger variance occurs for the sample having the larger number of observations the t becomes conservative with respect to control over P(EI). This occurs even for small variance differences. The most extreme combination of unequal n's and variances was represented in the case where $n_1 = 12$ and $n_2 = 36$ with $\sigma_1^2 = 1.0$ and $\sigma_2^2 = 40$ (VR = .025). Figure 1 presents the power curves for this situation (dashed lines). The loss of power for t is very noticeable. A less pronounced effect, with VR = .5 (solid lines) was found.

The situation where the large variance occurs for the sample having the fewer cases effects t by inflating P(EI). This effect was apparent with n-ratios as small as 1.25. A case in point is that in which $n_1 = 12$, $n_2 = 15$, as shown in Figure 2. When the variances were $\sigma_1^2 = 40.0$, and $\sigma_2^2 = 1.0$, The Welch, Scheffé, and Games and Klare procedures are nearly indistinguishable in terms of power. Each offers excellent control over P(EI) and while t is highly inflated. When the n-ratio is increased to 3.0 ($n_1 = 12$, $n_2 = 36$) the effect on t is extreme, with a highly inflated risk of a Type I error and a spuriously high power as shown in Figure 2. This effect is not mitigated when n's are increased to 24 and 72.

When n's diverge to the extent of 3:1 variance ratios as small as 2.0 ($\sigma_1^2 = 2.0$, $\sigma_2^2 = 1.0$) had a readily discernible effect upon t. Characteristically, the t revealed an inflated risk of a Type I error with correspondingly high power. Adequate control over P(EI) was attained by the other test procedures. The Welch solution had a slight edge in power over that of Scheffé. The Games and Klare method was superior to both the Welch and Scheffé methods in terms of power, but this was at the expense of an inflated P(EI) of .066.

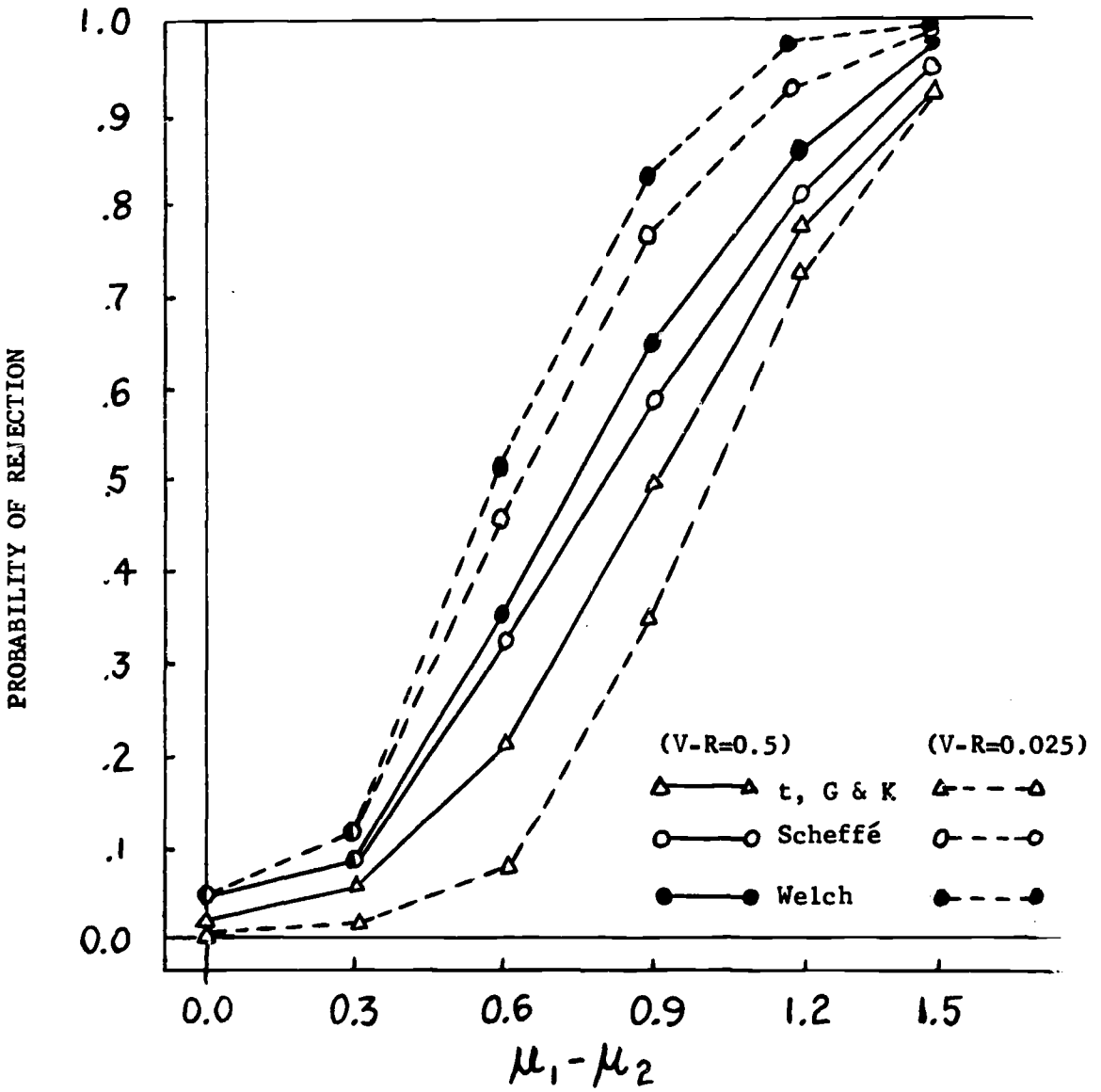

Figure 1. α = .05 power curves when large sample is drawn from the population with the large variance: Comparison of curves for slight vs. large variance inequality when n_1 = 12, n_2 = 36. Parameters: σ_1^2 = 1.0, σ_2^2 = 2.0 (solid lines) and σ_1^2 = 1.0, σ_2^2 = 40.0 (dashed lines).

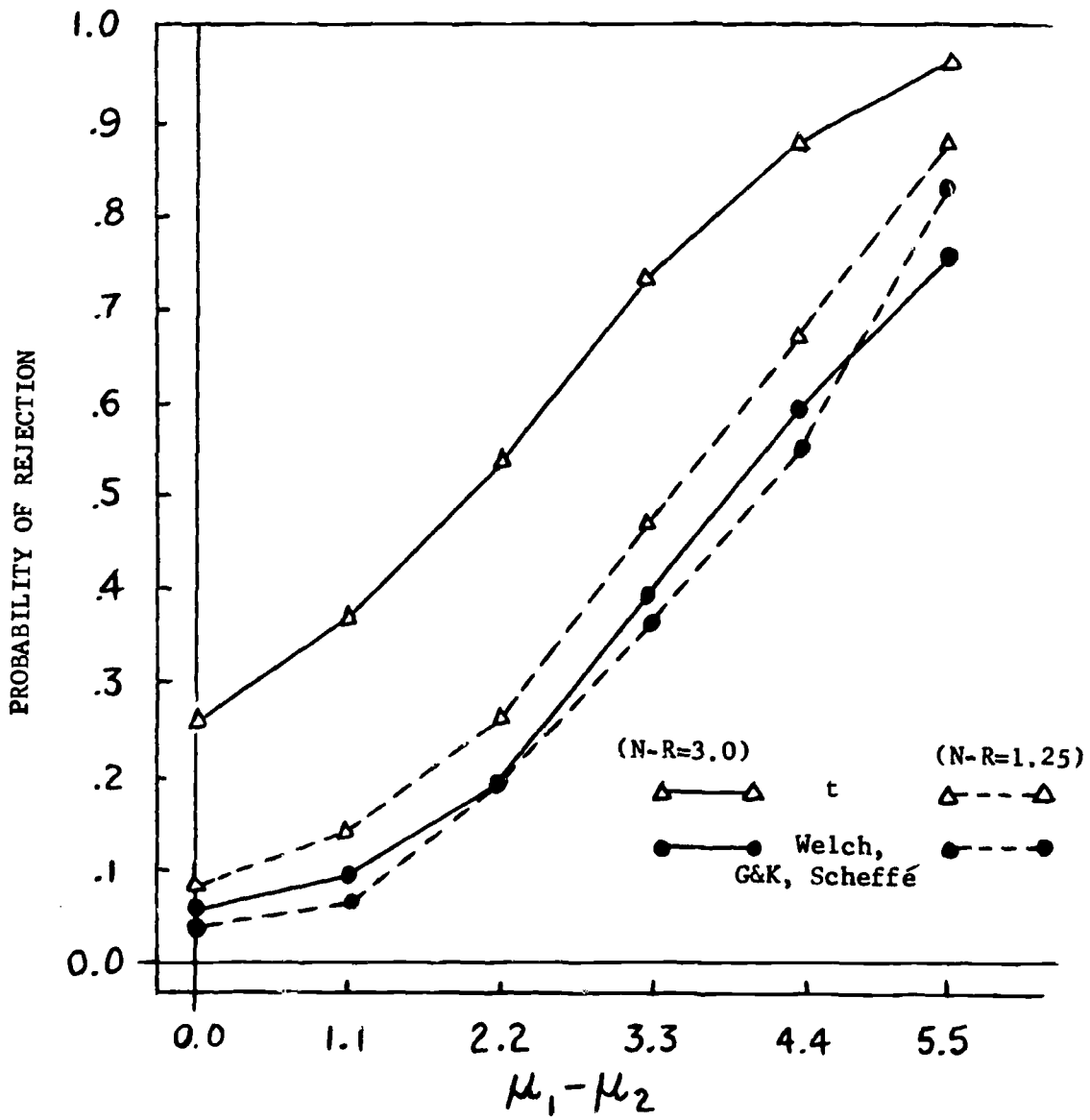

Figure 2, $\alpha = .05$ power curves when large sample is drawn from the population with the small variance: Comparison of curves for slight vs. large inequality of n's when $\sigma_1^2 = 40$, $\sigma_2^2 = 1.0$. Parameters: $n_1 = 12$, $n_2 = 15$ (dashed lines) and $n_1 = 12$, $n_2 = 36$, (solid lines).

DISCUSSION

If the population variances are known, the classical normal curve test (Blommers & Lindquist, 1960, p. 256) is the most powerful possible test and should be used. In practice, whenever $\underline{t}$ is considered, the population variances are unknown. Most of the time, there is little if any a priori basis to defend the assumption that the two unknown population variances are equal. The notion of additive treatment effects is a simplification that does not necessarily exist in reality. Treatment effects may be multiplicative with some individual difference parameter, or produce unequal variances because of some other interaction with subjects.

The present study demonstrates that the Games and Klare technique and the Scheffé test are superior to the conventional $\underline{t}$ test. However, the Scheffé test has the liability that it could produce different results when different randomizations are used. The Games and Klare procedure has the disadvantage of a multi-stage test. Nowhere in our study did either of these tests show any advantages over the $\underline{v}$-$\underline{W}$ technique. Wherever differences were noted, they were in the direction of the superiority of the $\underline{v}$-$\underline{W}$ test over all of its competitors.

The present study (Kohr, 1970) was conducted prior to the appearance of the two major theoretical papers supporting the use of the $\underline{v}$-$\underline{W}$ technique by Scheffé (1970) and Wang (1971). As such, we find ourselves in the embarrassing position of empirically demonstrating the superiority of a test that can now be shown to be superior by theoretical analyses.

The present state of knowledge suggests that the conventional $\underline{t}$ test be used only when $n_1 - n_2 = n$ and n is moderate to large. The $\underline{t}$ is tolerated in this case only because its permissive bias (when $\sigma_1^2 \neq \sigma_2^2$) is very mild. Even in this case, the user would be better off using the $\underline{v}$-$\underline{W}$ solution. However, when $n_1 = n_2$, then $\underline{v} = \underline{t}$, and the Welch critical value can vary only between

$t(\alpha, n-1)$ and $t(\alpha, 2n - 2)$. Using $\alpha = .05$, when $n = 21$, the critical value could vary only between 2.086 and 2.021. In this situation, the t test may be tolerated as a good approximation to the more accurate v-W solution, since the area between these two points is small. For small n's, the range of critical values possible under the Welch solution is larger, and the complete v-W solution should be used. When n's are unequal, the $v \neq t$, and the v-W solution should always be used. The Wang (1971) study considered a minimum sample size of 5, and the present study used a minimum size of 4. It is possible that with samples of only 2 or 3 cases, the v-W solution may not adequately control P(EI), but it is hard to conceive of any behavioral study that should be conducted with samples of this size.

REFERENCES

Aspin, A. A. Tables for use in comparisons whose accuracy involves two variances, separately estimated. Biometrika, 1949, 36, 293.

Bartlett, M. S. The information available in small samples. Proceedings of the Cambridge Philosophical Society, 1936, 32, 560-566.

Behrens, W. U. Ein beitrag zur fehlerberechnung bei wenigen beobachtungen. Landwirtschaftliche Jahrbucher, 1929, 68, 807-837.

Blommers, P. & Lindquist, E. F. Elementary Statistical Methods in Psychology and Education. Boston: Houghton-Mifflin, 1960.

Cochran, W. G. & Cox, G. M. Experimental Designs. New York: Wiley, 1950.

Dayton, C. M. The Design of Educational Experiments. New York: McGraw-Hill, 1970.

Dixon, W. J., & Massey, F. J. Introduction to Statistical Analysis. New York: McGraw-Hill, 1957.

Ferguson, G. A. Statistical Analysis in Psychology and Education. (2nd ed.) New York: McGraw-Hill, 1966.

Fisher, R. A. The fiducial argument in statistical inference. Annals of Eugenics, 1935, 6, 91-398.

Fisher, R. A., & Yates, F. Statistical Tables for Biological, Agricultural and Medical Research. (6th ed.) New York: Hafner, 1963.

Games, P. A. & Klare, G. Elementary Statistics: Data Analysis for the Behavioral Sciences. New York: McGraw-Hill, 1967.

Glass, G. V. & Stanley, J. C. Statistical Methods in Education and Psychology. Englewood Cliffs, New Jersey: Prentice-Hall, 1970.

Guilford, J. P. Fundamental Statistics in Psychology and Education (4th ed.) New York: McGraw-Hill, 1965.

Hays, W. L. Statistics for Psychologists. New York: Holt, Rinehart, & Winston, 1963.

Jeffreys, H. Notes on the Behrens-Fisher formula. Annals of Eugenics, 1940, 10, 48-51.

Kohr, R. L. A comparison of statistical procedures for testing $\mu_1 = \mu_2$ with unequal n's and variances. Unpublished doctoral dissertation, The Pennsylvania State University, 1970.

Knoble, H. D. PRAND, pseudo-random number generator. University Park, Pa.: Computation Center, The Pennsylvania State University, 1969.

Kirk, R. E. Experimental Design: Procedures for the Behavioral Sciences. Belmont, Calif.: Brooks/Cole, 1968.

Lewis, P. A., Goodman, A. S., & Miller, J. M. A pseudo-random number generator for the system/360. I.B.M. System Journal, 1969, 8, 136-146.

Myers, J. L. Fundamentals of Experimental Design (2nd ed.) Boston: Allyn & Bacon, 1972.

Payne, W. H., Rabung, J. R., & Bogyo, T. P. Coding the Lehmer pseudo-random number generator. Communications of the Association for Computing Machinery, 1969, 12, 85-86.

Satterthwaite, F. E. An approximate distribution of estimates of variance components. Biometrics, 1946, 2, 110-114.

Savage, L. J. The foundations of statistics reconsidered. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press, 1961, 575-586.

Scheffé, H. On solutions of the Behrens-Fisher problem, based on the t distribution. Annals of Mathematical Statistics, 1943, 14, 35-44.

Scheffé, H. Practical solutions of the Behrens-Fisher problem. American Statistical Association Journal, 1970, 65, 1501-1508.

Sukhatme, P. V. On Fisher and Behrens' test of significance for the difference in means of two normal samples. Sankhya, 1938, 4, 39-48.

Wang, Y. Y. Probabilities of the Type I errors of the Welch tests for the Behrens-Fisher problem. American Statistical Association Journal, 1971, 66, 605-608.

Welch, B. L. The generalization of 'student's' problem when several different population variances are involved. Biometrika, 1947, 34, 28-35.

Welch, B. L. Further note on Mrs. Aspin's tables and on certain approximations to the tabled function. Biometrika, 1949, 36, 293-296.

Winer, B. J. Statistical Principles in Experimental Design. New York: McGraw-Hill, 1962.

Wyatt, W. W., & Bridges, C. M., Jr. Statistics for the Behavioral Sciences Boston: Heath & Co., 1967.